Statistics Corner

Fundamental Issues in Experimental Design

by Bert Gunter

from Quality Progress June 1996

The methods of statistical design of experiments (DOE) can bc complex. This complexity makes it seem that DOE is an arcane science, needed only for special circumstances and not for the warp and weft of engineering and scientific practice. As a result, DOE still languishes as a too-little used and appreciated discipline, even though these methods were developed more [ban 75 years ago by R.A. Fisher and his colleagues. This is a tragedy. One can only speculate at the progress that could have been made and the resources saved had DOE been adopted as the standard approach to experimentation.

Adopting DOE dots not require that annics of statisticians be marshaled. As 1 mentioned in April's column, the fundamental principles and methods are straightforward and can easily bc taught as part of basic science and math education.[1] Every scientist, engineer, and technologist—indeed, anyone wishing to interact with science and technology - should understand these fundamental principles.

I do not know why conventional science has been so resistant to DOE methods. Perhaps it is because DOE challenges fundamental scientific practices (one-variabie-a[-a-time experimentation, for example) or because science views empirical model building as fundamen[ally different and inferior to the mechanistic models of physics and engineering. Perhaps it might even be due to the poor job statisticians have done in communicating the powerful and elegant principles that arc the real essence of DOE.

Whatever the reason. ignoring DOE is a loss. DOE is. in essence, a careful ex-position of the scientific method. It is fundamentally concemed with how people learn in a complex. interactive, and noisy world—that is. in the real world. it addresses the basic issues of what people know and how they know it. Ignoring such issues inevitably invites confusion and inefficiency.

An elaboration and defense

In this and subsequent columns, I will elaborate on and defend the previous statements. In doing so, I will avoid the conventional presentation of DOE as a suite of parlor tricks that improve quality and productivity. Instead, I will present DOE as a logical and inevitable way to deal with the world as it is. Denying DOE's relevance denies basic realities and thus compromises people's understanding and [he efficiency with which [hey learn.

Although these columns will take on a more philosophical bent than usual, I hope that you will nevertheless find them useful. If you already know about, practice. or teach DOE. I hope that these columns will illuminate your knowledge and help your efforts to convince others [o usc it. If you do not use DOE, I hope [hat they will compel you to learn more about it.

Two scientific mantras

My concern here is with the scientific learning strategy—how to increase one's knowledge of scientific phenomena. But what, exactly, are scientific phenomena? More precisely, what distinguishes scientific phenomena from religion, art, philosophy. emotion, favorite frozen yogurt flavor, and everything else that is not science? Without engaging in a broad philosophical debate, I propose two general distinguishing characteristics:

- Science is concemed with *broadly observable phenomena*. By (his, 1 mean that given normal human sensory abilities and the proper equipment, anyone should be able [o observe what is being described.
- Science is concerned with *repeatable phenomena*. A unique event, no matter how compelling or no molter how many people observed it, is simply not science.

i know that these principles still leave gray areas. You might argue. for example, that personality is an unobservable construct, but legitimate science nevertheless characterizes and studies it through behavior. Or you might argue that the meteorite that put a crater in your back yard is certainly a unique event, but as onc among a class of similar unique *events*, it is certainly a phenomenon amenable to scientific study. Clearly, working out the details here falls into the hands of the philosophers-of science, and I do not presume any such expertise. So I hope you will agree that observability and repeatability arc, for the purposes here, the essence of scientific matters.

Having said that, one arrives at the underlying justification for DOE: Without statistical thinking, the concepts of observability and repeatability arc inherently *contradictory*. Anything that is observed is observed with variability. Indeed, not only does exactly the same event never happen in exactly the same way twice, but even repeated observations——[t is, measurements—of the same event vary. So what is meant when science demands that an observation be repeatable? How repeatable is repeatable? When is a result the same or different---confirmation or contradiction? These questions cannot be coherently addressed without the framework of statistical thinking.

Let mc elucidate with a few examples:

- Eureka! You think that you've just discovered a ncw way to make semiconductor chips that run cooler. Of course, following the observability principle, you've got to actually make some of these new chips and observe a temperature difference. But when you make more than one, you discover that they do not all run at exactly the same temperature. In fact, although most run cooler than the existing chips, some of the existing chips are cooler than some of the new ones. Do the chips run cooler?
- You've got a great new chocolate soufflé recipc. The friend who gave it to you says it's real easy and never fails. You try it. It fails! Being stout of heart and chocolate of mind. you try it again. Success! Since ii's chocolate, you decide a third trial is called for. Success again. Is the recipe foolproof? (For science's sake, I note that a recent article claims that chocolate has more than 800 separate constituents, which makes it probably the most complex substance in the human gustatory spectrum.)
- You arc creating a complex computer model 10 describe the flow of blood through small arteries. The fluid dynamic model requires the fluid friction

and elasticity of the arterial walls as input parameters (among others). Like all measured physical constants, these are known only within (he limits of measurement precision. How does this uncertainty propagate through your model and affect its performance and your conclusions? Is this uncertainty inherent in any measured physical constant? (Be wary: Textbook calculations don't always give exact answers. Just because the fuzziness isn't explicitly stated doesn't mean it's not there.)

What is the point of this litany? Simply, that all observations involve uncertainty. Therefore, to deal rationally with the world as it is, you must define what is meant by "same" or "different" when all observations are blurred by variability. Statistical thinking-DOE, in particular-is required to do this.

The logic underlying statistical thinking: A simple comparative experiment

So how does statistics resolve these issues? The underlying approach can be clearly shown by considering a simple comparative experiment. Since *my* purpose is to focus on the concepts, not to teach statistical methods, I will omit the mathematical details, In any case, I should add that there are actually alternative methods to this experiment and its analysis that depend on the framework that is chosen and the assumptions that are made (such as the classical sampling, Bayes, empirical Bayes, and decision theories). Think of this in terms of how different architects might approach a home's design. An environmentalist might design a house one way, an architect working for a large builder another way, and one who works for a cement manufacturers' association still a third way. All of the houses would be functional, but the results would differ greatly. Of course, statistical approaches don't vary quite that much, but you should be aware that there is no single right way to translate the concepts discussed here into specific methods.

Consider, then, the following comparison of two treatments. The treatments might be two different suppliers of a raw material, two different working schedules, two different drugs, two different teachers, or two different flowmeter designs. Along with the two treatments, there are one or more outcome variables

(cont. *on p. 109)*

that are used to measure the goodness of the treatments. Such measurements can be subjective (e.g., the taste of frozen yogurt on a l-to-5 scale), categorical (e.g., pass or fail), or a continuous measurement (e.g., time, temperature, voltage, or cost). The kind of measurement affects both the details of the design and the analysis, but these matters are not of concern here. To keep this example straightforward, assume that the outcome variable is a continuous measurement.

How should you design the experiment to determine which treatment gives the better outcome? Of course, you can always just test each treatment once and compare the two outcomes. But if you were to repeat the tests, each treatment would give results that are somewhat different than what occurred previously. In other words, you cannot assume that two treatments are different when their outcomes differ because outcomes will differ *even when a single treatment is continually repeated.* The same could be said for averages if you repeat the tests and compare the averages for the two treatments. That is, if you were to repeat the whole series, the next pair of averages would be different from the first pair.

Therein lies both the crux of, and the solution to, the dilemma:

> The *only logically coherent way of determining whether (and how) the treatments differ is to compare the* variation in results from repeated tests in which the treatments actually change *with the* variation in results from repeated tests in which the treatments don't change.

In practical terms, this means that you must use the variability that is seen *within* each treatment group as a baseline to judge the observed differences between the two groups. The entire decision-making edifice of statistics is built on this simple, common-sense foundation.

You might not actually test the two treatments simultaneously. For example, you might compare historical data on one treatment with new tests on the second treatment-but there are risks in doing this. For example, other factors could change that could also have an effect on the results; thus, you risk confounding the extraneous variation with differences due to the treatments. Sometimes there are ways to mitigate such possibilities, but, in all cases, the replication principle underlies the details.

The key idea behind replication is that you must repeat the change to compare *the results that occur when the treatment doesn't change* with *the results that occur when it does.* As obvious as this principle seems, it is often at least partially violated in practice. How often have you heard, "There's no reason to spend the time and effort to redo the tests. We did it once, and we've seen what happens."

To be fair, sometimes it is difficult or impossible to obtain the kind of replications required to put the comparison on firm footing. In fact, even when experimenters do replicate, they often do so in a way that compromises the integrity of the conclusion, which can be seen in the following example.

Replication, real and imagined

Suppose that you want to compare how fast two different weed killers decompose in soil. The outcome measurement is obtained as follows: The chemicals are mixed with several different types of soils in plastic trays. The trays are then placed outside at a test site. After 30 days, soil samples are taken from the trays and the amount of chemical remaining in the soil is determined.

How should this procedure be carried out to adhere to the replication principle? The key idea is that you would like the variability within the results for each chemical to be the *same as the* variability in the results between the two sets, except for the differences due to the different chemicals. That way you can be sure that any possible (statistical) change in variability between groups is due to the chemical and not something else. Let's see what this means in terms of the necessary experimental procedures:

Scenario 1: A large container of soil type A is mixed with chemical No. 1 and then poured into several trays. This procedure is then repeated with chemical No. 2. The same set of procedures is carried out for soil types B, C, and D. If 10 trays are mixed for each chemical and soil type, does this give four groups of IO replicates for each soil type as the baseline for comparison?

The answer is no. Replicates are supposed to have the same variability between the groups, except for the different treatments, but in this scenario, the two treatments were mixed in each soil at different times. Variability could have been introduced if the instrument used to mix the soil was different or used differently. if different people mixed the soils, or if the humidity conditions were different. Within each treatment, however, these conditions were fixed, since large batches were mixed all at once. Thus, the preparation variability between the chemicals could well be greater than the preparation variability within the chemicals. This means that any difference seen between the chemicals might be due to preparation factors instead of the different chemicals.

How can you adhere to the replication principle? Ideally, you would want to reduce the preparation variability between the chemicals so that it matches the variability within the chemicals, but that is impossible. One person cannot possibly mix each chemical with each soil type simultaneously. Thus, you have to make the preparation variability within the chemicals the same as the preparation variability between the chemicals. This can be done by one person preparing each tray separately, using the same preparation procedures and tools. In doing so, the same variation in mixing factors will result within each chemical as well as between the chemicals.

Of course, experts in these matters might say that the effect of possible variation in mixing is so small that the extra effort is unjustified (and they might well be right). But, strictly speaking, what this scenario describes is often called duplication, not real replication. Duplicates look like replicates, but they exhibit less variability than they should—and this can compromise the conclusions that are reached.

Scenario 2: Suppose that one person prepares all 40 of the soil trays separately, using the same preparation procedures and tools. After 30 days, he or she takes soil samples from the trays containing one weed killer to determine the amount of chemical remaining and then repeats this process for the trays containing the second weed killer. Do these data provide for a clear comparison?

Again, the answer is no because the measurement variability between the chemicals is greater than the measurement variability within the chemicals. That is, they are duplicates with respect to the measurement process, not the preparation process. You need to measure the results so that the variability due to extraneous measurement noise within

(cont. *on). I 11)*

each chemical and soil type is the same as the measurement variability between chemicals and soil types.

Scenario 3: Suppose that one person prepares all 40 of the soil trays separately, using the same preparation procedures and tools, and then randomly measures the amount of chemical remaining in the trays. Now do you have true replicates?

Surprisingly, the answer is still no because this experiment is omitting what is probably the largest outside source of variability that might affect the comparison. Do you know what it is?

The answer is that just one batch of the first chemical is being compared with just one batch of the second. The batch-to-batch variability might be the largest extraneous source of variability. Strictly speaking, no matter what procedures are used, all that can be concluded is that one particular batch of the weed killer differs from another particular batch. If you wish to compare the first chemical in general with the second chemical in general—which, of course, you do—you would have to replicate with several different batches of each chemical.

And therein lies the rub. Replication is often difficult and sometimes impossible to do because it severely compromises the amount of experimentation that can be done given limited time and resources. For example, if the second weed killer is an experimental batch being tested for the first time, there very likely are no other batches. Even if there are, they are made in entirely different circumstances with entirely different batch-to-batch variability characteristics than the standard (production mode) batches. In such cases, you have no choice but to violate the replication principle.

When the replication principle is violated, it is important to understand what the risks are and realize that the experimental conclusions are, at best, tentative, no matter how well buttressed they are with statistical legerdemain. For those familiar with some of my earlier tirades, this is yet another example of an analytic study with which enumerative methods alone cannot deal. Subject-matter expertise and judgment must also be exercised.

Is there a better way to deal with replication?

Replication seems to violate what experimenters are taught is sound experimental practice: Conduct your

(cont. on p. 113)

experiments in a way that minimizes experimental variability. Fortunately, [his is still sound advice, but the replication principle demands that the variability be minimized *uniformly* over the procedures. Reducing experimental variability within treatments while having it remain between treatments compromises [he integrity of the learning process. Different results judged 10 be different because of different treatments might only vary duc to extraneous experimental variability.

From this discussion, it might appear that experimenters must always conduct comparative experiments in a way that inflates within-treatment variability to the same level as between-treatment variability. Not only does this require extra time and effort, but it also makes it more difficult to draw clear conclusions. Fortunately, variability inflation is only one approach, probably the crudest, to performing experiments that arc consistent with the replication principle. A better approach for reducing variability uses blocking. 1 will talk about this in detail in my August column, but let me suggest one version of how it could woi k for the weed killer experiment.

In the mixing procedure, instead of mixing all 40 batches separately, you could mix two trays of each chemical and soil type at a time (i.e., mix eight trays at one time), with the same mixing instrument, operator, humidity conditions, and so forth. This would be repeatcd five limes in all [0 get the same total of 40 trays. When analyzing the results, you would [hen compare the differences among chemicals and soil types among the four pairs within the groups of eight, thereby using this smaller expel-imcn[al variability as the baseline for comparison. Formal analysis procedures should [hen be used to combine the results from these five overall replicates to draw the overall conclusion. One classical, numerical formal analysis procedure is analysis of variance, or ANOVA, wi(h which many readers are undoubtedly familiar. In previous columns, however, 1 have also shown how less formal graphical methods—such as hierarchical do[plots—can also bc used to engage people's innate pattern-recognition capabilities to accomplish the same end.

Onc important conclusion that can be drawn from this discussion is that the *design* of an experiment—the detailed way in which it is conducted-dircc(ly affects the conclusions that can be reached and the data analysis procedures that must be used to reach [hem. C.artful thought and planning should always be devoted IO the design of an experiment in order to proceed in a way that maximizes the usc of scarce resources, takes advantage of all available ways to reduce experimental variability, and, most important, provides results that can be coherently analyzed according to the replication principle. No data analysis, no matter how sophisticated, can rescue a badly designed experiment or produce reliable information when the replication principle has been violated. That is why statistical experimental design is so important and why it should become part of all experimenters' standard practice.

References

1. Bert Gunter, "Data Mining: Mother Lode or Fool's Gold'?" *Quality Progress*, April 1996, pp. I 13-118.

2. Bert Gunter's six-part series entitled "Graphical Methods and Principles for Data Analysis 11" ran in the April 1995 (pp. 103-105), June 1995 (pp. 88-95), August 1995 (pp. 141- 143), October 1995 (pp. 129-137), December 1995 (pp. I I 4- I I 5), and February 1996 (pp. 109-1 I 5) issues of *Quality Progress*.

Bert Gunter is a contributing editor of *Quality Progress*, a member of ASQC, and a statistical consultant. Any questions or comments about this column may be sent 10 him at P.O. Box 9. Hopewell, NJ 08525 or through e-mail: bgunter@njcc.com.

What did you think about Ibis article?

Quality Progress needs your feedback. On the postage-paid reader service card inserted toward the back of this magazine, please circle the number that corresponds with your opinion of the preceding article.

Excellent	Circle #313
Good	Circle f1314
Fair	Circle #315
Poor	Circle #316